Journal of Computer Science

Research Article

Multi-FOM: A Unified Multimodal Deep Learning Framework for Intelligent Pre-Harvest Papaya Classification and Crop Valorisation

Sovers Singh Bisht[1], Sanjeev Thakur[1] and Sanjeev Panwar[2]

[1]*Department of Computer Science and Engineering, ASET, Amity University, Noida, Uttar Pradesh, India*
[2]*Technical Coordination Division, Indian Council of Agricultural Research, New Delhi, India*

Article history
Received: 25-07-2026
Revised: 20-08-2026
Accepted: 25-08-2026

Corresponding Author:
Sovers Singh Bisht
Department of Computer Science and Engineering, ASET, Amity University, Noida, Uttar Pradesh, India
Email: soverssingh@gmail.com

Abstract: Intelligent pre-harvest crop assessment has become an essential component of precision agriculture, where accurate crop quality evaluation enables informed decision-making and sustainable crop management. However, existing multimodal frameworks often exhibit limited data diversity and insufficient cross-modal semantic interaction, reducing the reliability of intelligent crop valorisation. Address these limitations, this paper proposes a Multimodal Focal-Mamba Optimized Valorisation Model (Multi-FOM), a unified deep learning framework integrating papaya image data and IoT sensor data for intelligent pre-harvest crop analysis. Initially, the Artificial Diffusion-based Image Augmentation Module (ADM) generates diverse and semantically consistent papaya image samples to improve dataset variability and model generalization. Subsequently, the Unified Multimodal Crop Harmonization Module (UMCH) pre-processes and harmonizes heterogeneous multimodal inputs to generate consistent crop representations. The Focal-Mamba ExcelFormer Tabular Representation Network (FMETRNet) employs a dual-stream architecture to extract complementary visual and IoT sensor features. These representations are integrated through the Perceiver-Q Cross-Modal Fusion Network (PQCF-Net), which performs latent cross-modal alignment and query-guided semantic refinement to produce unified crop features. The fused representation is further processed by the Deep Crop Intelligence Learning Module (DCIL) to capture crop health, fruit quality, maturity, environmental responses, and nutrient information. Finally, the Alpha-SiLU Activated Blue-Eared Hedgehog Optimized Lite Transfer Learning Model (αS-BHOT) performs adaptive feature optimization and intelligent crop classification to generate comprehensive crop valorisation outcomes. Experimental results demonstrate 99.67% classification accuracy, precision of 99.53%, recall of 99.61%, 99.72% F1-score, and a minimum MSE of 0.0123, confirming Multi FOM as an effective decision-support framework for next-generation precision agriculture.

Keywords: Papaya Crop, Precision Agriculture, Multimodal Deep Learning, Internet of Things, Crop Intelligence, Crop Valorisation

Introduction

The rising need for sustainable agriculture and quality crop production has pushed the use of Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), computer vision, remote sensing and Internet of Things (IoT)-based sensing technologies for intelligent crop monitoring and management to the fore. At the pre-harvest stage, the accurate assessment of crop health, maturity, ripeness, yield potential, and harvest readiness is crucial in boosting agricultural productivity, consumption of resources, minimizing crop losses, and timely decision making.

Traditional crop assessment approaches are largely based on manual visual inspection and expertise, and are often labor-intensive, time-consuming, subjective, and inappropriate in large-scale agricultural settings. Hence, Intelligent data-driven techniques have drawn great attention to enable an automatic and reliable analysis of crops. In recent years, deep learning has been shown to be effective for papaya disease identification (using mobile imaging) (Godapitiya and Mandira, 2025), computer vision techniques with the identification of chemically ripened bananas and papayas in real-time (Adhithya and Kathirvelan, 2025) and multimodal sensing approaches that combine remote and onsite information to estimate the pre-harvest ripening stage of avocado using machine learning algorithms (Lopez-Ruiz et al., 2025).

These advancements highlight the growing importance of intelligent crop monitoring systems for supporting precision agriculture and improving pre-harvest crop management. The use of multi-source remote sensing data and machine learning has been used to pre-harvest forecast rainfed wheat yield by using satellite-derived data to assist the prediction under varying environmental conditions (Bourbour et al., 2025). Similarly, multi-sensor data fusion techniques can be used to fuse together diverse data from various farming sources to produce more reliable predictive performance for the estimation of crop yield for corn and cotton (Rafi et al., 2026). Google Earth Engine has also been used to create mechanistic growth models and pre-harvest yield forecasting of chickpea production, along with Sentinel-2 imagery and AgERA5 meteorological data (Perach et al., 2025). Representative papaya fruits at different pre-harvest maturity stages used for image-based crop analysis is shown in Figure 1.

In another research, artificial neural network (ANN) was used to estimate pre-harvest yield of mango using the leaf nutrient variability and its capability was demonstrated for crop productivity assessment (Alebidi et al., 2025). In addition, large scale crop classification models using Random Forest have been examined to enhance spatio-temporal crop mapping and classification across various spatial regions, and to analyze the factors affecting the transferability of the models (Xie et al., 2025).

Fig. 1: Illustration of the pre-harvest papaya intelligence

Additionally, advanced computer vision and spectral sensing techniques have been developed recently, enhancing the non-destructive crop quality assessment, disease diagnosis, ripeness detection, and field monitoring system. Machine learning coupled with portable Near-Infrared (NIR) spectroscopy has been used for kiwi ripeness classification for rapid and non-invasive fruit quality assessment for precision farming applications (Altieri et al., 2025). The automated date fruit detection has also been explored using deep learning-based object detection models such as YOLO and Faster R-CNN, which were found to be effective in detecting fruits in various orchard environments (Lipinski et al., 2025).

Furthermore, machine learning crop forecasting models have played a role in sustainable agricultural management by improving the accuracy and reliability of crop production forecasting, leveraging a wide range of agricultural data sources (Singh et al., 2025). The support of UAV-based datasets for the classification of intercropping durian and papaya has also promoted the development of intelligent vision-based monitoring systems that can enhance the accuracy of crop identification for intercropping systems in difficult production conditions (Ngo et al., 2025). Furthermore, with the development of lightweight deep learning frameworks, efficient fruit disease diagnosis has become possible on resource-constrained agricultural equipment, making it more practical to deploy them in the field in real-time (Iftikhar et al., 2026) and optimized ensemble learning techniques have been used to combine multiple predictive models to provide better classification performance, which has improved the non-destructive classification of avocado ripeness in real time (Tipauksorn et al., 2025). While automated citrus disease grading has shown significant improvements in recent years, Zhu et al. (2025) proposed a unified framework that further enhanced disease assessment through advanced multimodal learning techniques.

Main Contribution of the Research

Existing approaches to crop monitoring using multimodal data have been identified as limitations, this study proposes Multi FOM framework for intelligent pre-harvest crop valorisation of papaya based on image and IoT sensor data. The key achievements of this work can be summarized as follows:

- ❖ To address Modality Representation Saturation (MRS), UMCH and FMETR-Net use modality-specific feature representations to retain visual and environmental information. In deep multimodal models, modality representation saturation (MRS) reduces representation variety and cross-modal information utilization by overemphasizing prominent modality-specific characteristics

- PQCF-Net solves Cross-Modal Semantic Misalignment (CMSM) by using adaptive semantic alignment and cross-modal feature fusion to improve image-sensor correlation. Cross-Modal Semantic Misalignment (CMSM) causes feature representations from different modalities to be inconsistent, resulting in poor multimodal fusion and decision performance
- To tackle the Decision Feature Divergence (DFD) phenomenon, DCIL learns high-level crop intelligence features that offer more discriminative representations for crop understanding.
- To enhance intelligent pre-harvest crop valorisation, ($\alpha S - BHOT$) optimizes the learned feature space and the crop classification is robust

Therefore, the paper proposes a novel multimodal deep learning framework for intelligent pre-harvest crop valorisation by effectively integrating visual and environmental information. The novelty of Multi FOM lies not in introducing entirely new backbone architectures, but in the unified integration of advanced visual, tabular, and cross-modal learning components for multimodal crop intelligence. To the best of our knowledge, such a comprehensive framework for intelligent pre-harvest papaya valorization has not been previously reported. Moreover, the remainder of this paper is organized as follows.

While the individual backbone architectures are derived from established deep learning models, their adaptation and integration within Multi FOM are specifically designed to address multimodal crop intelligence challenges, including image–IoT fusion, modality harmonization, and intelligent pre-harvest papaya valorization. The contribution therefore lies in the application-specific framework design rather than the introduction of entirely new backbone architectures.

Literature Survey

This section presents recent reviews of research on fruit quality analysis, crop health monitoring, maturity assessment and pre-harvest prediction using deep learning, machine learning and multimodal sensing approaches. Multimodal learning has evolved from traditional early fusion and late fusion strategies toward more sophisticated intermediate and attention-based fusion mechanisms. Early fusion combines heterogeneous features before model learning, whereas late fusion integrates decisions from modality-specific models. More recently, cross-attention and transformer-based approaches have enabled dynamic interaction between heterogeneous modalities, providing a strong foundation for modern multimodal intelligence frameworks such as the proposed Multi FOM architecture.

Deep Learning Based Fruit Quality, Maturity, and Disease Assessment

Nikam et al. (2025) developed a predictive quality grading and maturity detection method for dragon fruit using fused features of deep CNN models (VGG16 and Inception-ResNet-v2) and ensemble learning (Random Forest, Gradient Boosting and AdaBoost). The feature representation integrated with the system helped in discriminating quality and maturity attributes with accuracy 99.99% for quality grading and 99.69% for maturity detection by using the Random Forest and AdaBoost classifiers respectively. In addition to this, Revathi and Agash (2025) introduced a pomegranate disease diagnosis framework with DenseNet121, Grad-CAM++, Healthy-Based Deviation Scoring (HBDS), and a Retrieval-Augmented Generation (RAG)-based language model to provide treatment suggestions, with the classification accuracy of 99.35% and accurate estimation of pomegranate disease severity. To achieve high accuracy, Rayed et al. (2025) created MangoLeafXNet by combining a custom convolutional neural network with explainable AI methods, such as Grad-CAM, Saliency Maps and LIME, which achieved a 99.80% accuracy rate on various public datasets, with recall levels of 99.62%, precision of 99.50% and an F1-score of 99.56%. To classify the papaya freshness, Sar et al. (2025) implemented a PapayaFreshNet system that utilizes ResNet50, Transformer encoding, and Squeeze-and-Excitation attention mechanism which achieved 99.65% training accuracy and 97.68% testing accuracy. Rani et al. (2025) utilized curriculum learning with YOLOv5 and YOLOv7 models for the classification of pomegranate growth stage using a multi-source image dataset with the mean Average Precision (mAP) of 92.2%, precision of 90.1%, recall of 82.3% and an F1-score of 86.1%. In multi-class fruit ripeness detection, Kamat et al. (2025) found that YOLOv6 outperforms other models, achieving 99.5% mAP in real-time inference at 85.2 FPS, highlighting the suitability of object detection models for automated fruit quality assessment. Agricultural intelligence has also proliferated to intelligent sensing and recognition of diseases and even to pre-harvest prediction in multimodal learning approaches.

Multimodal Crop Monitoring and Pre-Harvest Prediction

Recent advancements have further extended agricultural intelligence toward multimodal sensing, crop monitoring, and pre-harvest prediction by integrating visual information with environmental and temporal data. Similarly, Tapia-Mendez et al. (2025) evaluated twelve object detection models of the MM Detection framework for the automated quality assessment of fruits and vegetables using a custom annotated dataset with multiple quality categories. DINO and DDQ performed the best among the evaluated models, with a mean Average Precision (mAP) of 0.65, and low training loss, indicating reliable quality recognition capabilities. Moreover, Srinivasan et al. (2025) established a dual-branch attention-guided vision network.

(DBA-ViNet) that combines global and local features and Grad-CAM for explainable fruit disease classification, which achieved the best performance with 99.51% accuracy, 99.61% recall, 99.30% precision, and 99.45% F1-score. To address the challenge of global pre-harvest forecast of crop yields, Zhuang et al. (2026) designed a diagnostic spatiotemporal multimodal fusion network, which automatically selects temporal and spatial learning modules based on the characteristics of the data set to obtain the lowest pooled NRMSE, with competitive MAPE and KGE values obtained across various countries and lead times. Wang et al. (2025) combined UAV time-series imagery with meteorological data to predict the yields of sugar beet, using a framework known as Stacked Long Short-Term Memory (Stacked-LSTM), which yielded R^2 values of 0.761 for sugar content, 0.531 for sugar root yield, and 0.478 for sugar yield. The findings in these studies illustrate the broad applicability and significant contributions of deep learning, multimodal sensing, and intelligent forecasting to agricultural monitoring and crop prediction. In addition, Rahman et al. (2026) proposed FocalNet, an attention-free architecture with fast multi-scale feature extraction and explainability for agricultural image analysis, including crop disease detection and fruit quality evaluation. Furthermore, Hatamizadeh and Kautz (2025) proposed MambaVision, a hybrid Mamba-Transformer architecture that extracts features efficiently and accurately for agricultural image processing applications including crop disease diagnosis and fruit quality evaluation. Moreover, the Transformer-based ExcelFormer model by Jha et al. (2026) incorporates complex feature interactions and may be used for crop production prediction, disease severity assessment, and precision farming. Similarly, Gorishniy et al. (2025) presented TabM, a parameter-efficient ensemble learning framework for tabular data, to enhance crop production predictions, soil analysis, and precision farming. Likewise, Khaliq et al. (2025) developed a hybrid Vision Transformer Perceiver IO architecture for reliable picture categorisation that may be used for crop disease diagnosis, plant health monitoring, and fruit quality evaluation. Additionally, Azad et al. (2025) presented HierarQ, a hierarchical Q-Former architecture for efficient video interpretation, which may be used for intelligent video analysis in crop growth monitoring, insect identification, and precision farming. Likewise, Nguyen et al. (2025) developed a superior YOLO-based plant disease detection model with αSiLU activation function, enhancing accuracy and resilience for smart agriculture applications such crop health monitoring and precision farming. Building upon these advancements, Dinler et al. (2025) developed the Hybrid Optimization Strategy for robust global optimisation, which improves agricultural AI feature selection, parameter optimisation, and crop management decision-making.

Research Gaps

A reliable pre-harvest papaya assessment needs to be a thorough understanding of the appearance of the crop and environmental conditions, which can be used to determine the quality, maturity, health of the crop and its readiness for cultivation. Most current multimodal crop assessment techniques, however, focus on modality-specific representation learning by applying a single feature extraction strategy to represent various visual features and heterogeneous sensor measurements before multimodal integration, which may lead to the Modality Representation Saturation (MRS) phenomenon. Thus, the local visual details, global context and complex environmental attributes are not well preserved, which results in incomplete modality-specific feature representation prior to multimodal fusion.

In addition, most multimodal crop assessment frameworks rely on the assumption that the features extracted from images and sensors align themselves during feature fusion. This assumption often results in the Cross-Modal Semantic Misalignment (CMSM) phenomenon, in which the characteristics of the crops in the image and the environment's responses are semantically inconsistent even if the features are aggregated correctly. Therefore, the intrinsic relationships between appearance and cultivation conditions of the crops are not modelled sufficiently, which causes the multimodal feature fusion to be less effective, and reduces the reliability of the unified crop intelligence learning.

Moreover, traditional crop prediction systems based on optimization rely mostly on numbers and accuracy of the predictions, without focusing on optimizing the discriminative feature space relevant to real agricultural decisions. This constraint results in the Decision Feature Divergence (DFD) phenomenon, which is that the optimum feature representation is not necessarily the best feature representation for intelligent valorisation of crops. Thus, the quality assessment, the prediction of harvest readiness, yield estimation, market value estimation, and personalised farmer recommendations could be less reliable.

Multimodal Focal-Mamba Optimized Valorisation Model (Multi-FOM)

Effective intelligent pre-harvest crop valorisation, effective integration of visual crop characteristics and environmental information is required to accurately assess crop quality, maturity, and disease condition and cultivation status. However, these different modalities are processed separately, with the limited interaction between them and low prediction accuracy. To address these limitations, the proposed Multimodal Focal-Mamba

DOI: 10.3844/jcssp.2026.2860.2878

Optimized Valorisation Model (Multi-FOM) aims to provide an end-to-end framework that combines feature representation in both streams, cross-modal feature fusion, deep crop intelligence learning, and intelligent crop valorisation. Figure 2 shows the overall architecture of the proposed framework.

As illustrated in Figure 2, the proposed framework consists of six sequential modules. Initially, the Artificial Diffusion-based Image Augmentation Module (ADM) generates augmented papaya images to improve dataset diversity. Subsequently, the Unified Multimodal Crop Harmonization Module (UMCH) performs preprocessing, normalization, and multimodal harmonization to papaya images and IoT sensor data to create unified crop representations. The Focal-Mamba ExcelFormer Tabular Representation Network (FMETR-Net) uses a dual-stream feature representation method, which uses FocalNet and Vision Mamba to extract rich visual feature representations from papaya images, and uses ExcelFormer and TabMamba to learn discriminative feature representations from the sensor measurement tab. Finally, the multimodal features extracted are combined using Perceiver-Q Cross-Modal Fusion Network (PQCF-Net), based on Perceiver IO for latent cross-modal alignment and Query-guided Multimodal refinement with the Q-Former, to generate a Unified Crop Feature Representation. This unified representation is then learned by the Deep Crop Intelligence Learning Module (DCIL) to provide higher level crop intelligence features, providing a description of the crop health, fruit quality, growth patterns, crop maturity, environmental responses, and nutrient-related characteristics. Lastly, the alphaSiLU activated Blue-Eared Hedgehog Optimised Lite Transfer Learning Model ($\alpha S - BHOT$) optimizes the features adaptively and intelligently classifies the crop to generate comprehensive results on crop valorisation, such as crop quality assessment, crop health assessment, harvest readiness prediction, yield prediction, market value estimation and personalised farmer recommendations, enabling intelligent pre-harvest crop management and sustainable precision agriculture. To mitigate the previously defined MRS and CMSM challenges, the proposed UMCH and PQCF-Net modules perform multimodal harmonization and semantic alignment before unified feature learning. Subsequently, DCIL and Alpha-SiLU-based Transfer Learning Block enhance crop intelligence extraction and adaptive decision optimization.

The proposed hierarchical fusion strategy was designed to address the complementary characteristics of visual and IoT modalities. Specifically, FocalNet and Vision Mamba capture local spatial patterns and long-range contextual dependencies from crop images, whereas ExcelFormer and TabM learn heterogeneous sensor relationships from tabular environmental data. Perceiver IO subsequently transforms these heterogeneous representations into a compact latent space, reducing modality-specific redundancy while preserving salient information.

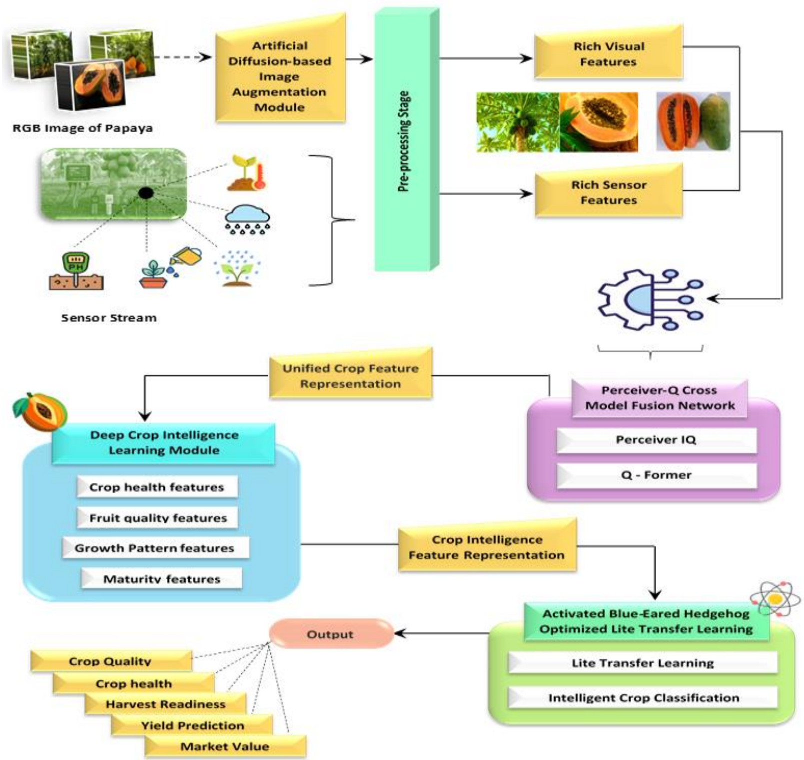


Fig. 2: Structural view of Multimodal Focal-Mamba Optimized Valorisation Model

DOI: 10.3844/jcssp.2026.2860.2878

Finally, Q-Former facilitates query-guided cross-modal interaction learning, enabling the generation of unified crop intelligence representations. This progressive fusion mechanism was adopted to improve multimodal representation quality and decision robustness in complex agricultural environments.

Artificial Diffusion-Based Image Augmentation Module (ADM)

The ADM serves as the initial stage of the proposed Multi-FOM framework, aiming to improve the diversity and generalization capability of the papaya image dataset before multimodal preprocessing. Images of papaya collected in the actual agricultural environments may be corrupted by illumination changes, viewpoint changes, background complexity, scale differences, and environmental disturbances, which affects the robustness of feature learning when using the original images. To overcome this, ADM developed a diffusion-based image generation strategy that enables the synthesis of realistic and semantically consistent papaya image variations while retaining crucial visual features like papaya texture, color distribution, disease symptoms, maturity patterns, and structural integrity. The proposed ADM can be trained using these images to capture the underlying image distribution and can then generate a variety of samples with biological meaning, without changing the inherent characteristics of the crops themselves, unlike traditional augmentation methods based on single geometric or photometric transformations. As a result, the augmented papaya image dataset offers more visual information, improves the learning ability of later deep feature extraction modules, decreases overfitting, and bolsters the overall robustness of intelligent pre-harvest crop classification under a wide range of agricultural environments. The enhanced papaya images are then sent to the UMCH for pre-processing, normalization and multi-modal data harmonization.

Unified Multimodal Crop Harmonization Module (UMCH)

The UMCH serves as the second stage of the framework. It is fed with the enhanced papaya dataset created by the ADM along with the IoT sensor readings to produce a unified and coherent multimodal representation for further deep feature learning. The dimensionality, sampling frequency, feature distributions, and noise characteristics of image data and sensor observations are different due to the different acquisition modalities. Because of these multiple differences in the inputs, modality imbalance, information redundancy, and feature distribution inconsistencies can arise, lowering the performance of multimodal representation learning. Data Harmonization is done using Image samples and IoT sensor records were integrated through a common Image_Name identifier. Only matched records were retained for multimodal learning. As the dataset was pre-curated, no additional missing-value imputation or timestamp synchronization procedures were required. Sensor attributes were subsequently normalized using Min-Max scaling. Hence, UMCH systematically conducts pre-processing, normalization, synchronization, and harmonization of the enhanced papaya images with the IoT sensor data to satisfy the visual properties of the papaya fruits and the environmental data obtained from the IoT sensors. The multimodal crop dataset is presented as two complementary information sources: Augmented Papaya image dataset and IoT sensor dataset. Suppose the whole multimodal crop data set is given by Equation (1):

$$D_M = \{I, S\} \tag{1}$$

Where M is the full multimodal crop dataset that consists of two complementary data sources. The set of papaya images is given as $I = \{I_1, I_2, \dots, I_n\}$ and the IoT sensor observations are given as $S = \{S_1, S_2, \dots, S_n\}$. Additionally, n is the number of all the multimodal crop samples to be considered for synchronization. The image data set mainly consists of visual data representing the appearance of the crop, fruit texture, color distribution, disease symptoms and maturity characteristics. In the case of image acquisition in the practical agricultural environment, however, there will be illumination variations, background noise, camera noise, and inconsistencies in scale. These variations are reduced by pre-processing each image by downsampling, intensity normalisation and pixel value standardisation before passing it to the feature extraction stage. Normalized image representation is represented by Equation 2:

$$\hat{I}_i = \frac{I_i - \mu_I}{\sigma_I} \tag{2}$$

Here I_i is the original papaya image, μ_I is the mean of the intensity of the image, and σ_I is the standard deviation of the distribution of the pixel's intensity. After image standardization, $\hat{I}_i$ is the normalized papaya image, and is used for visual feature representation learning. At the same time, the information from the IoT sensors is complementary to the environment measured with temperature, humidity, soil moisture, pH, concentration of nutrients and other cultivation parameters. These variables are measured in different units of physical measurement and have different ranges of numbers and may bias the learning process when integrated with visual information. Therefore, sensor data is first cleaned for missing or inconsistent data and then is scaled to a uniform numerical range. The normalized sensor vector is calculated as Equation 3:

$$\hat{S}_i^j = \frac{S_i^j - S_{min}^j}{S_{max}^j - S_{min}^j} \tag{3}$$

DOI: 10.3844/jcssp.2026.2860.2878

Here S_i^j denotes the j^{th} sensor attribute of the i^{th} crop sample. The first term S_{min}^j denotes the minimal value of the respective sensor attribute for the entire data set, and the second term S_{max}^j the maximal value. Thus, the normalized value of the sensor, $\hat{S}_i^j$ allow all the environmental measurements to be represented in a same numerical range, and thus consistent multimodal learning. Both modalities undergo preprocessing individually, and are then temporally synchronized and harmonized, ensuring a one-to-one correspondence between visual observations and environmental measurements. This harmonisation process creates a consistent, multimodal representation, retaining complementary information from both sensing modalities, while preserving structural consistency of the synchronised crop samples. The harmonized crop representation is defined as Equation 4:

$$H_i = \Psi(\hat{I}_i, \hat{S}_i) \tag{4}$$

Here,$\Psi(\cdot)$ is the proposed multimodal harmonization function used for the integration of the normalized image representation $\hat{I}_i$ and the normalized sensor representation $\hat{S}_i$. The output H_i is then the harmonized sample of the crops, which will retain both the visual properties and the environment information that are complementary for the following multimodal feature representation learning. The UMCH module's output is therefore a harmonised multi-modal crop dataset comprising of standardised image representations and synchronised environmental sensors' measurements. This one representation helps to reduce the modality differences and removes the superfluous information while maintaining complementary features for effective multimodal feature learning. Despite its harmonized representation of the crop dataset, the low-level visual and environmental observations contained in the dataset are still too limited to express complex crop semantics. Hence, the multimodal dataset H is directly sent to the (FMETR-Net) for receiving discriminative visual and tabular representations of features, which are learned independently from image and IoT sensor modalities respectively and then multimodal feature fusion is performed in the next stage.

Focal Mamba Excel Former Tabular Representation Network (FMETR-Net)

After the multimodal crop harmonization process conducted by the UMCH, the resulting multimodal crop data is provided to the proposed FMETR-Net to learn the discriminative features for the multi-modal crops. Previously mentioned in research gap, the current multimodal crop assessment techniques are plagued with the Modality Representation Saturation (MRS) problem that results in limited representations of the features in the image and sensor information before multimodal fusion. To tackle this drawback, FMETR-Net takes a dual-stream learning approach, with FocalNet-Vision Mamba learning rich visual representation and ExcelFormer-TabM learning informative sensor representation. The feature representations generated are complementary to the modality-specific information and are then passed on for cross-modal fusion. First, the harmonized multimodal crop data obtained by UMCH are presented in two forms that are the harmonized papaya image data, and the harmonized IoT sensor data:

$$D_H = \{I_H, S_H\} \tag{5}$$

D_H is the harmonized multimodal crop data from the UMCH module, I_H denotes the harmonized papaya image data, and S_H denotes the harmonized IoT sensor data. Each of these modalities is represented and processed independently in parallel branches of representation learning before being fused by FMETR-Net, to maintain modality-specific information.

Image Representation Learning

The image representation stream aims to use the hierarchical context learning ability of FocalNet and the global dependency modelling ability of Vision Mamba to extract discriminative visual features from harmonized papaya images. First, the image patches of each harmonized papaya image are projected into an embedding space of the latent images to create an initial image representation. This embedding process converts the raw image pixels into smaller, more meaningful feature vectors that help deep learning algorithms understand the image:

$$Z_0 = f_e(I_H) \tag{6}$$

Where as $f_e(I_H)$ is an image patch embedding function and Z_0 is the initial image feature representation after embedding the harmonized papaya image. This basic visual information is the basis for further context learning at higher hierarchical levels. After the embedding of images, FocalNet conduct hierarchical contextual learning through depth-wise convolution and nonlinear activation functions to layer up the local structural information at different contextual levels. Instead of self-attention methods which calculate full token-to-token interactions, FocalNet effectively leverages contextual information at different receptive fields for extracting fine-grained papaya attributes like surface appearance, disease symptoms, papaya shape, fruit texture, and color transitions, which are correlated with the level of maturity:

$$Z_l = f_a(GeLU(DWConv(Z_{l-1}))) \tag{7}$$

Here $DWConv(\cdot)$ is the depth-wise convolution operation that captures local spatial dependence, $GeLU(\cdot)$ is the Gaussian Error Linear Unit activation function to add nonlinear feature transformation, $f_a(\cdot)$ is the hierarchical contextual aggregation function, and Z_l is the contextual feature representation generated at the l^{th} hierarchical abstraction level. The proposed FMETR-Net can be effectively used to preserve the multi-scale visual information and improve the discriminative local feature extraction through this hierarchical learning strategy. The number of contextual feature maps created in the process of learning in a hierarchical system is different for each system, but the contribution of each of them depends on the visual information in papaya image. To enable this, FocalNet uses an adaptive gated aggregation approach that is able to selectively mix contextual information from various hierarchical levels into a common feature representation. This mechanism strengthens visual responses that are most informative of the context and reduces irrelevant or less discriminative visual responses:

$$Z_F = \sum_{l=1}^{L+1} G_l \odot Z_l \tag{8}$$

Where G_l denotes the learnable gating weight associated with the l^{th} contextual level, Z_l represents the contextual feature map extracted from that level, and $\odot$ indicates element-wise multiplication between the gating weights and the corresponding contextual features. The aggregated contextual representation is denoted by Z_F which integrates informative local features obtained from multiple hierarchical levels into a unified contextual representation. Consequently, this adaptive aggregation process enables FMETR-Net to preserve the most discriminative visual characteristics required for accurate pre-harvest papaya analysis while minimizing the influence of redundant contextual information.

Local Visual Feature Representation

The unified contextual representation is then converted to a discriminative local visual representation through the focal modulation mechanism. This process uses the original image features embedded in the image and the aggregated contextual information to highlight the most informative visual features while maintaining the fine-grained spatial details. Therefore, the proposed FMETR-Net can well learn local representations related to papaya surface texture, surface colour variation, disease symptoms and maturity-related feature:

$$F_{Local} = q(Z_0) \odot h(Z_F) \tag{9}$$

Where $q(\cdot)$ denotes the query projection function that projects the embedded image features into the feature interaction space, $h(\cdot)$ represents the linear transformation applied to the aggregated contextual representation, Z_0 corresponds to the initial embedded image representation obtained from the patch embedding layer, and Z_F denotes the aggregated contextual representation generated through the gated hierarchical contextual learning process. The resulting F_{Local} represents the discriminative local visual feature representation learned by the FocalNet module. Pre-harvest crop assessments also need to model long-range contextual relationships across the whole papaya fruit, beyond the effective capture of fine-grained local visual characteristics by the FocalNet module. Fruits often have important information about their maturity, health status and structural uniformity in distant image regions, which is communicated through spatial interactions. Thus, the obtained local visual representation is then passed to the Vision Mamba module, where it uses a selective state space modelling mechanism to effectively learn global contextual relationships and with linear complexity. From the hidden state evolution in the Vision Mamba architecture follows:

$$h_t = \bar{A} h_{t-1} \bar{B} F_{Local} \tag{10}$$

The hidden state at the current step of the sequence is denoted as h_t the hidden state at the prior step of the sequence is denoted as h_{t-1}, the state transition matrix responsible for propagating the information across the sequence is denoted as $\bar{A}$, and the input projection matrix that combines the local visual representation F_{Local} into the state-space model is denoted as $\bar{B}$. The recursive state transition allows Vision Mamba to learn long-range semantic relationships in the entire image efficiently. The learning contextual information is then converted into a global feature representation that encodes the characteristics of the structure of the various regions of the papaya fruit. This is not just a local feature extraction operation, and the network can interpret the overall visual composition and appearance of the fruit:

$$F_{Global} = \bar{C} h_t \tag{11}$$

Where $\bar{C}$ is the output projection matrix, which maps the hidden state to the global contextual feature representation, and the global semantic representation learned by the Vision Mamba module is F_{Global}. This representation maintains the long-range contextual information which is represented by the overall fruit structure, maturity distribution and global visual consistency. Additionally, Vision Mamba incorporates the selective state-space modelling and convolutional feature enhancement techniques to enhance the learned representation further. Through the selective scan operation, sequential information is obtained, being combined with convolutional spatial information to form a complete visual depiction, with local texture information and global contextual dependency being simultaneously captured:

$$F_V = Linear\left(Concat\left(Scan(F_{Global}), Conv(F_{Global})\right)\right) \quad (12)$$

The selective scan operation $Scan(\cdot)$ is used to extract long-range sequential dependencies, the convolution operation $Conv(\cdot)$ is used to preserve local spatial information, the feature concatenation operation $Concat(\cdot)$ is used to concatenate sequential and convolutional representations, and the linear projection layer $Linear(\cdot)$ is used to project features to the final embedding space. The resulting F_V is the Rich Visual Feature Representation that combines hierarchical local features extracted by FocalNet with global contextual information learned by Vision Mamba. The image representation stream of the proposed FMETR-Net therefore yields an extensive visual feature representation that allows for a good characterisation of papaya fruit texture, colour distribution, disease symptoms, maturity state and general structural features. The generated sections are then directly passed to the PQCF-Net and then adaptive cross-modal interaction, and multimodal feature fusion are applied to generate a unified crop feature representation.

Perceiver-Q Cross Modal Fusion Network (PQCF-Net)

Rich feature representations of visual and sensor modalities are extracted after feature representation learning using the FMETR-Net without relying on the image modality and IoT sensor modalities. However, as pointed out in the research gap, direct fusion of these heterogeneous representations can result in the phenomenon of Cross-Modal Semantic Misalignment (CMSM), which is the semantic inconsistency of complementary visual and environmental information. To tackle this challenge, the proposed PQCF-Net introduces an adaptive cross-modal alignment mechanism based on Perceiver IO and Q-Former to build meaningful semantic interaction prior to multimodal fusion. As a result, a common representation of the crop features is derived for intelligent pre-harvest crop analysis. Initially, the rich visual and rich sensor feature representations generated by FMETR-Net are jointly considered as the multimodal input of PQCF-Net:

$$F_{IN} = \{F_V, F_S\} \quad (13)$$

For the multimodal feature input in the proposed PQCF-Net, F_{IN} denotes the multimodal feature input. F_V denotes the Rich Visual Feature Representation extracted from the FocalNet Vision Mamba image stream. F_S denotes the Rich Sensor Feature Representation learned from the sensor stream of the ExcelFormer-TabM. These complementary feature representations describe the visual appearance of papaya fruits as well as their environmental sensing characteristics. The visual and sensor features are extracted from different networks; they have different embedding distributions and feature dimensions. Thus, two feature representations are separately linearly projected into a common latent embedding space before performing cross-modal interaction. This is a modality projection that reduces the modality difference while retaining the discriminative information obtained from each modality:

$$F'_V = \frac{W_V F_V + b_V}{F'_S W_S F_S + b_S} \quad (14)$$

Here, W_V and W_S are the learnable projection matrices from the visual stream and sensor stream, respectively, and b_V and b_S are their respective bias vectors. The feature representations of the projected feature F'_V and F'_S share a common embedding dimension, which helps to efficiently interact with each other during the fusion process. After feature projection, then both modalities are combined into a single multimodal feature sequence, which is processed by the Perceiver IO module. While traditional self-attention models have quadratic complexity in the input sequence length, Perceiver IO achieves feature interaction via a small latent representation that allows it to efficiently learn long-range interactions across different modalities while maintaining a low computational cost:

$$F_M = Concat(F'_V, F'_S) \quad (15)$$

Where $Concat(\cdot)$ represents the feature concatenation operation to merge the projected visual and sensor representations into a single multimodal feature sequence. F_M is the combined multimodal representation prior to latent feature encoding. The concatenated multimodal representation is then fed to the Perceiver IO module where latent cross-modal alignment is performed. In this process, the learnable latent vectors interact iteratively with the visual, and sensor features to learn their complementary semantic relationships while filtering redundant modality-specific information. This leads to a compact latent representation of heterogeneous multimodal features that preserves important crop characteristics:

$$F_L = PerceiverIO(F_M) \quad (16)$$

Here, $PerceiverIO(\cdot)$ refers to the cross-modal encoding function performed by the Perceiver IO module, while the F_L denotes the latent multimodal representation of the features derived from heterogeneous visual and sensor information, which are aligned in a shared latent feature space. Perceiver IO effectively integrates multimodal information into a common latent representation, but there are some semantic relationships that are still underrepresented for each modality. PQCF-

Net further includes the Q-Former module to further improve multimodal interactions by learnable query embeddings. The Q-Former is designed to selectively sample the most discriminative multimodal features from the latent representation by using a small number of trainable query vectors, which boosts the discrimination and semantic consistency of the features. The learnable query embeddings are represented as:

$$Q = \{q_1, q_2, \dots, q_n\} \tag{17}$$

Where Q denotes the set of learnable query embeddings, q_i represents the i^{th} query vector, and n corresponds to the total number of learnable queries employed for multimodal feature refinement. Using the learnable query embeddings, the latent multimodal representation is refined through the Q-Former module according to $F_R = QFormer(F_L, Q)$. Finally, the refined multimodal representation is projected to a linear projection layer, which is a unified crop feature representation. This operation combines complementary information learned from both modalities in a single high-level feature embedding which can be used for subsequent learning of crop intelligence:

$$F_U = Linear(F_R) \tag{18}$$

The linear projection operation is denoted as $Linear(\cdot)$, the refined multimodal feature representation is denoted as F_R, and the final Unified Crop Feature Representation is denoted as F_U, which is obtained by the PQCF-Net. Thereby, PQCF-Net is able to seamlessly fuse complementary visual and environmental sensing information by means of a latent cross-modal alignment module and a query guided multimodal refinement module. The resulting Unified Crop Feature Representation F_U contains all detailed information about the papaya appearance, maturity, disease symptoms, environmental conditions and crop growth characteristics. This representation is then sent to the DCIL where higher-level crop intelligence features are learnt for intelligent pre-harvest crop valorisation.

Deep Crop Intelligence Learning Module (DCIL)

The DCIL is receiving the Unified Crop Feature Representation from the PQCF-Net. The integrated multimodal features are then further processed and high-level crop intelligence information is learnt by exploring complex relations between visual crop features and environmental sensing features. Instead of multimodal fusion, DCIL aims at transforming the integrated feature representation to a more discriminative feature space for capturing the overall crop condition. The aggregated crop feature representation derived from PQCF-Net be represented as F_U.The deep crop intelligence learning process is presented as:

$$F_C = \phi(F_U) \tag{19}$$

Where F_U is the Unified Crop Feature Representation generated in the previous stage, $\phi(\cdot)$ is a deep learning Crop Intelligence learning function, and F_C is a learned Crop Intelligence Feature Representation. In this transformation, the joined multimodal features are transformed to a smaller representation that preserves meaningful crop intelligence information. The learned crop intelligence representation is then enhanced by adding high level semantic information related to crop growth and field conditions. In this way, the framework can capture complementary agronomic information necessary for intelligent crop assessment:

$$F_{CI} = Enhance(F_C) \tag{20}$$

Here F_C denotes the initial Crop Intelligence Feature Representation, F_{CI} denotes the Crop Intelligence Feature Representation after the feature enhancement operation $Enhance(\cdot)$. An improved representation is possible, as it can learn the underlying relationships between the multimodal features integrated together to capture various crop health-related features, fruit quality information, crop growth, maturity, environmental responses, and nutrient-related features. Lastly, the improved crop intelligence representation is passed to the optimization stage to enable intelligent crop decision-making and multi-objective crop valorisation:

$$F_{OPT} = F_{CI} \tag{21}$$

Where F_{OPT} denotes the feature input forwarded in an optimized manner to the next optimization stage, and F_{CI} is the final Crop Intelligence Feature Representation outputted by the module. The DCIL converts the integrated representation into a detailed crop intelligence representation, which is a representation that includes high-level semantic information about the crop health characteristics, fruit quality, growth patterns, maturity status, environmental responses and nutrient related characteristics. The resulting Crop Intelligence Feature Representation (F_{OPT}) is then passed on to the $(\alpha S - BHOT)$ where the features are optimized and intelligent crop classed.

Alpha-SiLU Activated Blue-Eared Hedgehog Optimized Lite Transfer Learning Model ($\alpha S - BHOT$)

The proposed $\alpha S - BHOT$ is the last step in the framework and is adaptively optimized before classification by the Crop Intelligence Feature Representation obtained by the DCIL. Conventional optimization techniques frequently have a challenge known as Decision Feature Divergence (DFD), where the features optimized numerically cannot be the most discriminative features in the process of intelligent crop valorisation, as discussed in the research gap. To address this issue, $\alpha S - BHOT$ combines Alpha-SiLU-based Transfer Learning Block activation with Blue-Eared Hedgehog Optimization and a light-weight transfer learning backbone to maximize the learned multimodal features while eliminating redundant information. The optimized feature representation is then applied to correctly classify crop in pre-harvest stage and valorize it:

$$F_O = \text{argmax}\Omega(F_{CI};\theta) \quad (22)$$

F_{CI} is the Crop Intelligence Feature Representation obtained by the DCIL module, while $\Omega(\cdot)$ is the proposed optimization function, and θ is the optimization parameters learned during the optimization process, and F_O represents the optimized crop feature representation. With this optimization strategy, the proposed framework improves the quality of the learned multimodal features by highlighting the very discriminative information while discarding the redundant feature components. After feature optimization, the optimized feature representation is sent to the light weighted transfer learning classifier, where the intelligence crop classification takes place. The proposed Alpha-SiLU-based Transfer Learning Block mechanism is used in the classifier to enhance the nonlinear feature learning and to enhance the separability of different crop conditions. The optimized feature representation allows the classifier to effectively classify different crop features by ensuring strong feature discrimination and efficient knowledge transfer. The classification process is represented by:

$$P(y|F_O) = \sigma_\alpha(WF_O + b) \quad (23)$$

Where $P(y|F_O)$ denotes the probability distribution of the predicted crop classes, W represents the learnable weight matrix of the lightweight transfer learning classifier, b denotes the corresponding bias vector, F_O is the optimized crop feature representation, and $\sigma_\alpha(\cdot)$ represents the proposed Alpha-SiLU activation function. The resulting probability distribution reflects the confidence of the model for each possible crop condition based on the optimized multimodal features. At the last stage, comprehensive pre-harvest crop valorisation outcomes are created using the predicted crop class with the highest classification probability. These outcomes deliver intelligent decision-support information for precision agriculture, by combining the appearance of the crop, the environmental sensing information and learned crop intelligence characteristics into an integrated decision-making framework. The last valorisation step of the crop is formulated as:

$$V = \psi(argmaxP(y|F_O) \quad (24)$$

Where $\psi(\cdot)$ denotes the crop valorisation function, $P(y|F_O)$ identifies the predicted crop class with the highest classification probability, and V represents the final crop valorisation output. The decision support outputs produced are crop quality assessment, crop health assessment, harvest readiness prediction, yield prediction, market value estimation and personalized crop recommendations for farmers, which helps farmers make accurate and informed pre-harvest management decisions. Thus, the proposed Alpha-SiLU Activated Blue-Eared Hedgehog Optimised Lite Transfer Learning Model $\alpha S - BHOT$ efficiently transforms the learned crop intelligence representation to the optimized classification results and complete crop valorisation results. The proposed framework offers an efficient and reliable decision-support system for intelligent pre-harvest papaya analysis and sustainable precision agriculture, making it a valuable tool for this application.

Algorithm 1: Multimodal Focal-Mamba Optimized Valorisation Model (Multi-FOM)

***Input*:** Papaya RGB Images $\boldsymbol{I_{img}}$, IoT Sensor Measurements $\boldsymbol{S_{sensor}}$
***Output*:** Crop Quality CQ, Crop Health CH, Harvest Readiness HR, Yield Prediction YP, Market Value MV, Farmer Recommendation FR

Begin
1: ***Acquire*** multimodal crop samples $\boldsymbol{\{I_{img}, S_{sensor}\}}$
2: ***Initialize*** ADM,UMCH, FMETR-Net, PQCF-Net, DCIL and $\boldsymbol{(\alpha S - BHOT)}$
3: for each multimodal sample do
4: Generate augmented papaya image $\boldsymbol{I_{aug}}$ using ADM
5: ***Normalize*** $\boldsymbol{I_{aug}}$ and preprocess $\boldsymbol{S_{sensor}}$
6: ***Harmonize*** augmented image and sensor representations using UMCH
7: Split harmonized data into image and sensor streams
8: ***Extract*** image embeddings from $\boldsymbol{I_{aug}}$
9: Learn local contextual features using FocalNet
10: Learn global visual dependencies using Vision Mamba
11: ***Generate*** Rich Visual Feature Representation (RVF)
12: Learn sensor representations using ExcelFormer and TabM
13: ***Generate*** Rich Sensor Feature Representation (RSF)
14: Project RVF and RSF into a common latent space
15: ***Perform*** latent cross-modal alignment using Perceiver IO
16: ***Refine*** multimodal features using Q-Former queries
17: ***Obtain*** Unified Crop Feature Representation (UCF)
18: Learn Crop Intelligence Features using (DCIL)
19: ***while*** optimization criterion is not satisfied do

20: ***Optimize*** intelligence features using **($\alpha S - BHOT$)**
21: end while
22: ***if*** optimized feature representation is obtained then
23: ***Classify*** crop using Lite Transfer Learning classifier
24: ***else***
25: Re-optimize intelligence features
26: ***end if***
27: ***Generate*** {CQ, CH, HR, YP, MV, FR}
28: ***end for***
29: Return comprehensive crop valorisation outcomes
End

Results and Discussion

The following section presents the experimental evaluation of the proposed Multi FOM framework using multimodal papaya image and IoT sensor datasets. The obtained results are analysed to evaluate multimodal representation learning, cross-modal feature fusion, optimization effectiveness, and overall prediction performance for intelligent pre-harvest crop valorisation.

Dataset Description

Further, the proposed Multi FOM framework leverages a common multimodal information dataset (Dari, 2025), which includes papaya RGB images and IoT sensor measurements, for an intelligent valorisation of papaya crops before harvest. The data set includes a total of 428 images of papaya fruits infected with the four disease conditions: Anthracnose (100 images), Black Spot (100 images), Healthy (128 images), and Phytophthora (100 images), with the agro-climatic conditions of Tamil Nadu, India. To improve the diversity of the training data and enhance the generalization capability of the proposed framework, five data augmentation techniques were applied to each original image, namely Random Rotation (±20°), Random Zoom In, Random Zoom Out, Horizontal Flip, and Brightness & Contrast Adjustment. As a result, five augmented versions were generated for every original image, increasing the dataset size from 428 to 2,568 images. Consequently, the final dataset contains 600 Anthracnose images, 600 Black Spot images, 768 Healthy images, and 600 Phytophthora images. To ensure systematic data organisation, each image is coupled to image metadata like the image name, disease category, crop variety, and region of cultivation. In parallel with the visual modality, the IoT sensor data set includes environmental, and soil related measurements synchronized with the images, such as temperature, humidity, soil moisture, soil pH, nitrogen, phosphorus, potassium, light intensity, atmospheric CO_2 concentration, and wind speed. The image and the records from the sensor are associated by the same Image Name identifier making the correspondence between the visual observation and environment conditions is one-to-one. Sensor features were normalized using Min-Max scaling, while categorical state information was encoded using one-hot encoding. Image and IoT records were matched through a common Image_Name identifier, and stratified sampling was applied to create training (72%), validation (8%), and testing (20%) subsets with a fixed random seed of 42 for reproducibility. This multi-modal information fusion allows simultaneous learning of the appearance of the crop and the cultivation attributes of it, which in turn facilitates comprehensive pre-harvest analyses. Meanwhile, the dataset is used to test the ability of the proposed framework to accurately predict other aspects of crop quality such as health status, harvest readiness, yield, market value and personalized crop recommendations in a variety of agricultural scenarios.

Experimental Configuration

In this section, the experimental setup that was used to test the proposed Multi-FOM framework is presented. Moreover, the multimodal dataset, Image Augmentation strategy, the preprocessing strategy, the hyperparameters of the networks, the training configuration, the metrics used for the evaluation, and the environment in which it was implemented, using the python programming language.

The merged multimodal dataset was partitioned using stratified sampling to maintain class balance, with 80% allocated for training and 20% for testing. Subsequently, 10% of the training data was reserved for validation, resulting in an overall distribution of approximately 72% training, 8% validation, and 20% testing. A fixed random seed of 42 was used to ensure reproducibility. The proposed Multi-FOM framework was implemented using Python, PyTorch, OpenCV, Scikit-learn, NumPy, and Pandas. Training was performed with a batch size of 32, learning rate of 0.001, and 20 epochs using the Adam optimizer, Cross-Entropy Loss, and Cosine Annealing learning-rate scheduler. Experiments were conducted on a workstation equipped with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA RTX 4060 GPU (8 GB VRAM) running Windows 11 with CUDA-enabled PyTorch support.

Simulation Assessment of Multi-FOM

Then, moving on, the simulation results provided in this section show a stable learning process in multi-modal learning, efficient feature optimization and reliable crop intelligence generation under different pre-harvest agricultural conditions.

Figure 3 shows the individual data augmentation techniques applied to a sample image of a healthy papaya to increase the dataset diversity and improve the robustness of the deep learning model. It shows how various augmentation procedures manage geometric and photometric fluctuations while keeping class label. By showing the model from different angles, rotation and horizontal flipping increase orientation invariance.

DOI: 10.3844/jcssp.2026.2860.2878

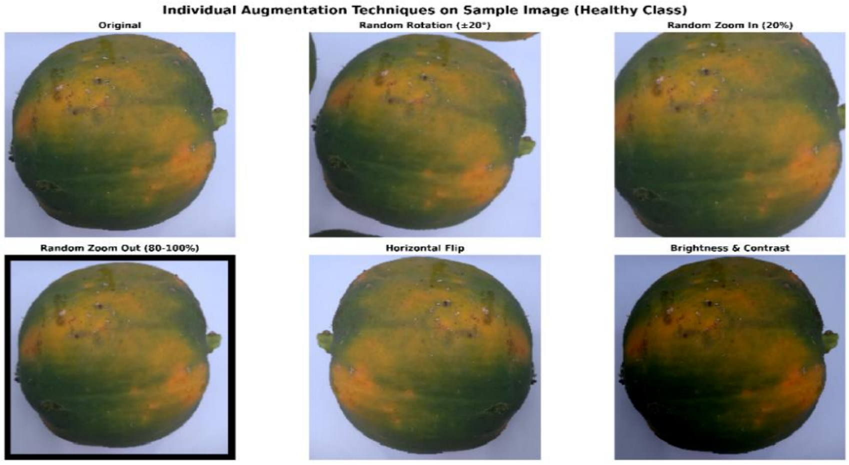


Fig. 3: Illustration of Individual Data Augmentation Techniques for Healthy Papaya Image

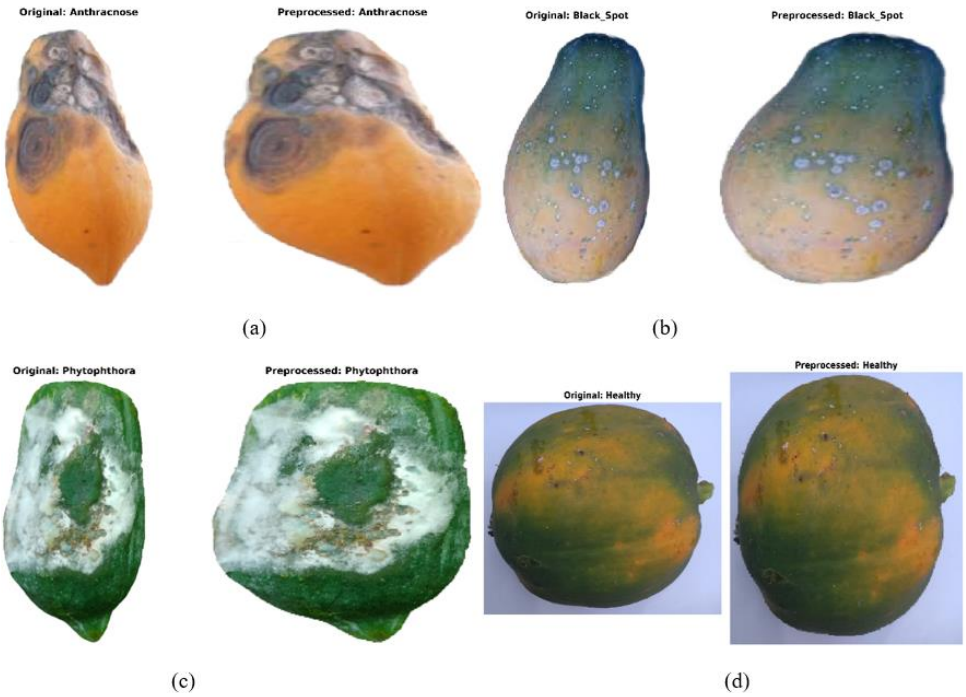


Fig. 4: (a-d): Visual demonstration of (a) anthracnose, (b) black spot (c), phytophthora-infected and, (d) normal papaya images

Zoom-in actions highlight local lesion textures and disease patterns, whereas zoom-out modifications preserve fruit structure. Adjusting brightness and contrast simulates field acquisition lighting conditions. These augmentations boost feature variety, minimise overfitting, and makes the visual branch of the proposed Multi-FOM architecture more robust and generalisable. The transformations are random rotation (±20°), random zoom-in (20%), random zoom-out (80-100%), horizontal flip and brightness & contrast adjustment, where the class label is unchanged, but the orientation, scale and illumination are changed in a controlled manner.

The pre-processing step of simulation involves modification of the original papaya image both in terms of resizing and standardizing it as in Figure 4 (a), in order to improve the visual lesion features for robust two-stream feature extraction. Then, Figure 4 (b) aims to normalize the input image size and enhance the structural blight features, improving the visual input to the feature extraction stage utilizing Mamba. Meanwhile, Figure 4(c) signifies the reduced image dimensions and also strengthens pattern of blight structures for better processing in the next image feature extraction step using the Mamba model for the visual input. The healthy papaya

DOI: 10.3844/jcssp.2026.2860.2878

image in Figure 4 (d) is resized and optimized in spatial orientation for base feature extraction in learning with the dual stream network.

Further, on analysis, there are clear ranges within the normalized features as in Figure 5, and the feature with the highest median is ~0.60, while the maximum (~0.85) is humidity, followed by temperature (~0.52), and then the others. The median (0.38) and minimum (0.18) level of moisture in the soil is least. Nitrogen, potassium and soil pH have intermediate medians of 0.46 to 0.50.

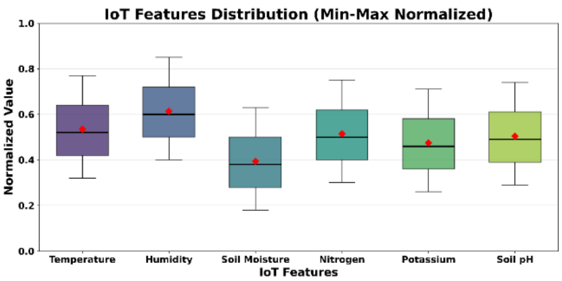


Fig. 5: Distribution of min-max normalized IoT sensor features

Moreover, the interaction of environmental data is evident with high correlation values as seen in Figure 6 (a), The correlation matrix in Figure 6(a) reveals strong interdependencies among several environmental and soil attributes. For instance, Rainfall exhibits high positive correlations with Soil Moisture (0.88) and Humidity (0.79), while Soil Moisture is also strongly correlated with Humidity (0.82). Similarly, Light Intensity shows a strong positive correlation with Temperature (0.73), whereas it exhibits inverse relationships with Humidity (-0.66) and Soil Moisture (-0.60). Soil nutrients demonstrate moderate cross-correlations, with the strongest association observed between Phosphorus and Potassium (0.55). These relationships indicate the presence of both complementary and redundant information among sensor variables. The Excel Former branch effectively captures such inter-feature dependencies through attention-based learning, enabling robust tabular feature representation while mitigating the impact of multicollinearity during multimodal fusion within the proposed Multi-FOM framework.

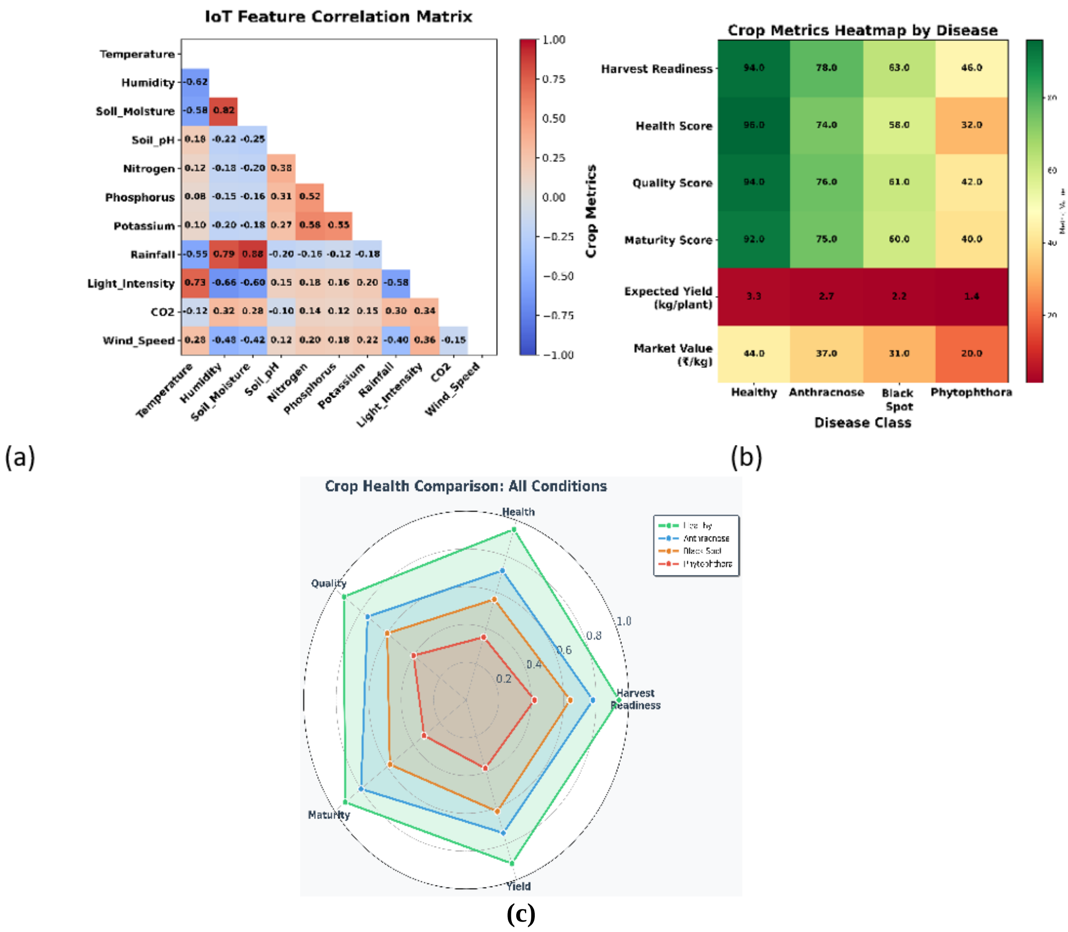


Fig. 6: (a-c): Simulation examination of (a) correlation matrix of environmental IoT Features, (b) heatmap matrix of varying disease severity classes, and (c) radar profile comparison of unified multi-task crop attributes

DOI: 10.3844/jcssp.2026.2860.2878

Moving on, the heatmap presented in Figure 6 (b) highlights substantial multi-task degradation under disease conditions. In this, under health, the best score is achieved by healthy crops (96.0), under harvest readiness (94.0) and under yield (3.3 kg/plant). However, in case of phytophthora disease, the performance is significantly reduced with health score of 32.0, quality score of 42.0, yield of 1.4 kg/plant and market value of ₹20.0/kg. Further, coming to radar chart in Figure 6 (c) signifies a set of metrics that are normalized to the conditions. Meanwhile, crops to be healthy it would be in an almost optimum profile, ~0.96 for health; ~0.94 for quality. Anthracnose infection is estimated to decrease values to ~0.74 (health) and ~0.78 (readiness). Black spot reduces performance to ~0.60 while Phytophthora severely affects performance in drastic ways; health is reduced to ~0.32 and quality to ~0.4.

Performance Evaluation of Multi-FOM

In this, coming onto performance assessment, which demonstrates the Multi FOM's effectiveness for accomplishing strong prediction, enhanced learning efficiency, and uniformity of decision-making across multimodal datasets.

Initially, during performance, the Multi FOM model achieves precise convergence over 20 epochs as provided in Figure 7 (a). The training accuracy starts at ~58% and gradually increases until it reaches peak accuracy at epoch 19 of 99.5%. At the same time, the accuracy begins with ~25%, has a temporary drop to 75% at epoch 8, but rapidly recovers to 100% at epoch 10, defining good generalization. Further, on loss analysis offered in Figure 7 (b), where the convergence curves present strong optimization in 20 epochs. The training loss decreases from ~0.90 at epoch 0 to a very small ~0.05 at epoch 19. Loss for validation decreases rapidly from ~2.50 until minor convergence spikes are overcome at epoch 4 (~0.85) and epoch 8 (~0.65) and remain stable to less than 0.10 from epoch 10 onwards.

Finally in Figure 7 (c), the Multi-FOM framework provides excellent classification results with metrics better than 0.98 for every class. The accuracy of classified anthracnose is 0.995, and for healthy, the accuracy is 0.993 (precision) and 0.994 (recall). Meanwhile, the areas of high recall are 0.993 for black spot and 0.994 for phytophthora, with 0.986 precision recorded for both diseases.

To evaluate the contribution of individual components within the proposed Multi-FOM framework, as shown in Table 1, an ablation study was conducted by systematically removing key modules and measuring the resulting performance degradation. As shown in Table X, the complete Multi-FOM architecture achieved the highest classification accuracy of 99.67%, precision of 99.53%, recall of 99.61%, and F1-score of 99.72%. The removal of UMCH, Vision Mamba, PQCF-Net, DCIL, and αS-BHOT resulted in consistent reductions across all evaluation metrics, confirming the effectiveness of each component. Among the evaluated variants, the exclusion of αS-BHOT produced the largest performance drop, whereas the removal of PQCF-Net and DCIL significantly affected multimodal feature fusion and crop intelligence learning capabilities. These results demonstrate that each module contributes positively to the overall performance, robustness, and discriminative power of the proposed Multi-FOM framework.

Benchmark View of Multi-FOM

Comparative evaluation offered in this section further demonstrates that Multi-FOM achieves best results in the multimodal representation learning and intelligent crop assessment tasks for all tested modalities.

The very first comparison seen in Figure 8 (a), the proposed Multi-FOM framework achieves best performance in all the measures such as 99.6% accuracy, 99.4% precision, 99.5% recall and 99.6% F1-score.

(a) (b) (c)

Fig. 7: (a-c): Performance estimation of (a) training and validation accuracy, (b) training and validation loss, and (c) Multi-FOM model across varied papaya pathogenic classes

DOI: 10.3844/jcssp.2026.2860.2878

Table 1: Ablation study of the proposed Multi-FOM framework showing the contribution of individual modules to overall classification performance

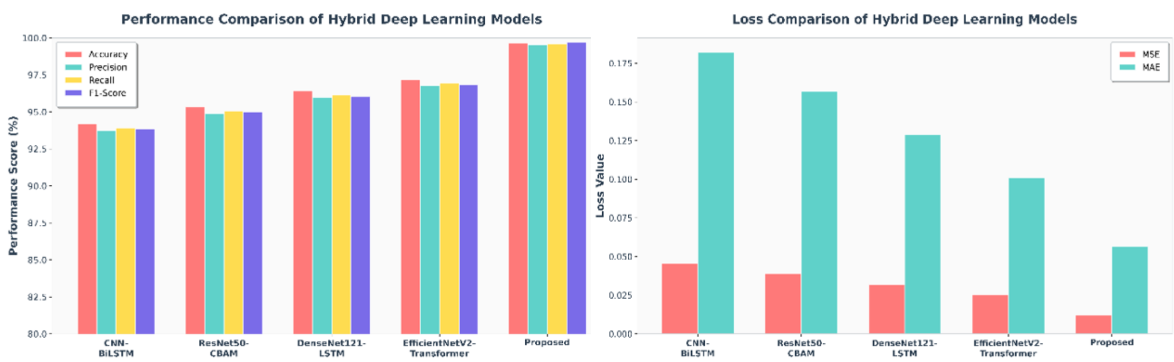
Model	Accuracy	Precision	Recall	F1
Full Multi-FOM	99.67	99.53	99.61	99.72
Without UMCH	98.91	98.84	98.79	98.81
Without Vision Mamba	98.73	98.68	98.61	98.64
Without PQCF-Net	98.46	98.39	98.32	98.35
Without DCIL	98.21	98.15	98.08	98.11
Without αS-BHOT	97.94	97.86	97.79	97.82

It achieves significantly higher accuracy (97.4%) compared to the competitive models CNN-BiLSTM (96.4%), DenseNet121-LSTM (96.4%), and ResNet50-CBAM (95.3%) along with the baseline CNN-LSTM structure with the lowest accuracy (~94.1%). Then, based on the errors obtained as provided in Figure 8 (b), the Multi-FOM framework attains the minimum error rate with the values of MSE as ~0.012 and MAE as ~0.056. It sharply outperforms the EfficientNetV2-Transformer (MSE ~0.025, MAE ~0.101) and DenseNet121-LSTM (MSE ~0.032, MAE ~0.128). The highest errors occur in ResNet50-CBAM (MAE ~0.156) and CNN-BiLSTM (MSE ~0.045, MAE ~0.182).

Discussion

During discussion, it is seen that the proposed Multi-FOM is a framework, which is scalable and multimodal, enabling the integration of visual and environmental knowledge, for reliable pre-harvest crop assessment. Its innovative architecture, by its results provided in Table 2 and 3 enables the deployment in precision agriculture, smart farming platforms, autonomous field monitoring, edge deployments in IoT solutions, digital agriculture ecosystems, and intelligent deployments in agricultural decision-support applications.

Moreover, as presented in Table 3, the proposed Multi-FOM achieved a classification accuracy of 99.67%, an F1-score of 99.72%, an MSE of 0.0123, and an MAE of 0.0567. Similarly, Table 2 shows that the proposed framework achieved 98% harvest readiness with an estimated market value of ₹45/kg for healthy papaya. Furthermore, the comparative analysis presented in Table 4 indicates that the proposed Multi FOM consistently outperformed the deep learning models reported in Rayed et al. (2025), achieving the highest accuracy, precision, recall, and F1-score. This performance improvement demonstrates the effectiveness of the proposed multimodal framework in providing more reliable and accurate pre-harvest crop assessment than existing image-based deep learning approaches.



(a) (b)

Fig. 8: (a-b): Comparative examination of Multi-FOM against (a) quantitative metrics, and (b) loss of hybrid DL models

Table 2: Intelligent pre-harvest crop valorisation estimation for different papaya disease classes

Disease Class	Harvest Readiness	Grade	Health	Risk	Yield	Market Value
Healthy	90-98%	A	Healthy	Low	3.0-3.5 kg	₹42-45/kg
Anthracnose	70-85%	B	Moderate	Medium	2.4-2.9 kg	₹34-39/kg
Blackspot	55-75%	C	Diseased	Medium	2.0-2.5 kg	₹28-34/kg
Phytophthora	30-55%	D	Critical	High	1.2-2.0 kg	₹18-25/kg

Table 3: Comparative examination of the proposed Multi-FOM Framework with Existing DL Models

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	MSE	MAE
CNN-BiLSTM	94.18	93.74	93.92	93.83	0.0456	0.1823
ResNet50-CBAM	95.36	94.91	95.08	94.99	0.0389	0.1567
DenseNet121-LSTM	96.41	95.98	96.15	96.06	0.0321	0.1289
EfficientNetV2-Transformer	97.18	96.76	96.93	96.84	0.0254	0.1012
Proposed (Multi-FOM)	99.67	99.53	99.61	99.72	0.0123	0.0567

Table 4: Benchmark Comparison of the Proposed Multi-FOM Rayed et al. (2025)

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet50	94.00	91.00	92.00	95.00
EfficientNetB2	93.00	90.00	93.00	92.00
XceptionNet	92.00	85.00	92.00	91.00
MobileNetV2	87.00	81.00	87.00	86.00
CNN	83.00	89.00	83.00	92.00
InceptionNetV3	91.00	91.00	91.00	90.00
DenseNet121	94.00	91.00	93.00	92.00
Proposed (Multi-FOM)	99.67	99.53	99.61	99.72

Conclusion

From this extensive research, as a whole DL model named Multi FOM was presented, which integrates RGB images and measurements from IoT sensors to accurately assess papaya crops that would improve the preparation for harvest. The proposed framework was systematically designed to mitigate the drawbacks of current multimodal crop analysis approaches by six sequential stages, namely Artificial Diffusion-based Image Augmentation, multimodal data harmonization, dual-stream visual and tabular feature representation learning, and adaptive cross-modal semantic fusion, deep crop intelligence learning, and optimized transfer learning-based classification. Moreover, through the incorporation of complementary visual attributes and environment observation information, the framework was able to build complete representations of crop intelligence to make trustworthy crop quality evaluation, monitoring and health management, crop harvest readiness prediction, crop yield estimation, crop market value assessment and farmer recommendation generation. The proposed architecture improved the feature interaction among multimodal sensors while maintaining sufficient discriminative crop information for a strong support to agricultural decision making in various cultivation environments. The proposed framework is evaluated through experiments and compared with various hybrid DL models, demonstrating its superior classification and prediction capabilities in all instances. The proposed Multi-FOM showed that it achieved a classification accuracy of 99.67%, precision of 99.53%, recall of 99.61% and an F1-score of 99.72% capability to value crops before harvest, a clear result of its efficient performance that it can be considered reliable and accurate. The outcomes show that the proposed framework serves as an efficient, intelligent and scalable decision support system that can promote precision agriculture by enabling relevant multimodal crop intelligence and farm sustainable management.

Future Scope

As when it comes to future, Multi FOM can be combined with foundation vision-language models, digital twins, edge intelligence, federated learning, hyperspectral imaging, UAV-assisted monitoring and self-supervised multimodal learning. By doing this way, Multi FOM allow for scalable, privacy-preserving, explainable, and autonomous crop intelligence in next-generation precision agriculture ecosystems. Future research will combine reinforcement learning with adaptive multimodal reasoning for dynamic crop management and autonomous agricultural decision-making. Intelligent agricultural valorisation systems can be more transparent, reliable, and sustainable using explainable AI and climate-aware predictive analytics. Further research will focus on lightweight model optimization, compression, pruning, quantization, and edge deployment evaluation to improve real-time applicability and computational efficiency in precision agriculture.

Acknowledgment

The authors would like to sincerely thank Amity University, Noida for providing the facilities, tools, and learning environment required to successfully complete this study.

Funding Information

There was no funding.

Author’s Contributions

Sovers Singh Bisht: Conceptualization, research design, experimentation, manuscript writing, and preparation of the original draft.

Sanjeev Thakur: Data analysis, validation, and use of analytical tools.

Sanjeev Panwar: Research coordination and supervision.

Ethics

The authors declare that there are no ethical issues associated with this research. They remain committed to addressing any ethical concerns that may arise following the publication of this manuscript.

References

Adhithya, S., & Kathirvelan, J. (2025). Computer vision-based detection and classification of chemically ripened bananas and papayas at vendor site through deep learning AI models using real-time dataset. *Results in Engineering*, *26*, 104730. https://doi.org/10.1016/j.rineng.2025.104730

DOI: 10.3844/jcssp.2026.2860.2878

Alebidi, A., Almutairi, K. F., Al-Obeed, R. S., Makhasha, E., Aboukarima, A. M., & Abdel-Sattar, M. (2025). Pre-harvest mango yield prediction using artificial neural networks based on leaf nutrient variability. *PeerJ*, *13*, e20013. https://doi.org/10.7717/peerj.20013

Altieri, G., Laveglia, S., Rashvand, M., Genovese, F., Matera, A., Mininni, A. N., Calabritto, M., & Renzo, G. C. D. (2025). Portable NIR Spectroscopy Combined with Machine Learning for Kiwi Ripeness Classification: An Approach to Precision Farming. *Applied Sciences*, *15*(11), 6233. https://doi.org/10.3390/app15116233

Azad, S., Vineet, V., & Rawat, Y. S. (2025). HierarQ: Task-Aware Hierarchical Q-Former for Enhanced Video Understanding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8545–8556. https://doi.org/10.1109/cvpr52734.2025.00799

Bourbour, H., Abdolahipour, M., Abdollahi, H., Abiar, E., & Mashal, M. (2025). Pre-harvest forecasting of rainfed wheat yield in Iran using multi-source remote sensing and machine learning. *Agricultural Water Management*, *322*, 110005. https://doi.org/10.1016/j.agwat.2025.110005

Dari, A. (2025). Indian hybrid papaya multimodal dataset. *Kaggle*. https://www.kaggle.com/datasets/ajithdari/indian-hybrid-papaya-multimodal-dataset/data

Dinler, Ö. B., Bektemyssova, G., Ahmed, M. A., Ibraheem, I. K., Smerat, A., Montazeri, Z., Dehghani, M., Malik, O. P., Şahin, C. B., AL-Salih, A. A. M. M., & Eguchi, K. (2025). Blue-eared Hedgehog Optimization (BEHO): A Nature-inspired Metaheuristic for Robust and Efficient Global Optimization. *International Journal of Intelligent Engineering and Systems*, *18*(11), 133–148. https://doi.org/10.22266/ijies2025.1231.08

Godapitiya, M. V. N., & Mandira, T. K. N. (2025). A Deep Learning Approach for Papaya Disease, Pest and Maturity Identification via Mobile Imaging. *2025 7th International Conference on Advancements in Computing (ICAC)*, 1–6. https://doi.org/10.1109/icac69156.2025.11361477

Gorishniy, Y., Kotelnikov, A., & Babenko, A. (2025). Tabm: Advancing tabular deep learning with parameterefficient ensembling. *International Conference on Learning Representations*, 77899–77935.

Hatamizadeh, A., & Kautz, J. (2025). MambaVision: A Hybrid Mamba-Transformer Vision Backbone. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 25261–25270. https://doi.org/10.1109/cvpr52734.2025.02352

Iftikhar, S., Khattak, H. A., Saadat, A., Ameer, Z., & Zakarya, M. (2026). Efficient fruit disease diagnosis on resource-constrained agriculture devices. *Journal of the Saudi Society of Agricultural Sciences*, *25*(1), 19. https://doi.org/10.1007/s44447-025-00094-9

Jha, A. N., Hoehn, M., Mazin, A., Gyawali, N., Aziz, H. M. A., Caragea, D., & Fitzsimmons, E. J. (2026). Crash Severity Prediction Using Transformer-Based ExcelFormer Model Using Kansas Crash Data. In *International Conference on Transportation and Development 2026* (pp. 60–71). https://doi.org/10.1061/9780784487013.006

Kamat, P., Gite, S., Chandekar, H., Dlima, L., & Pradhan, B. (2025). Multi-class fruit ripeness detection using YOLO and SSD object detection models. *Discover Applied Sciences*, *7*(9), 931. https://doi.org/10.1007/s42452-025-07617-7

Khaliq, A., Ahmad, F., Rehman, H. U., Alanazi, S. A., Haleem, H., Junaid, K., & Andrikopoulou, E. (2025). RETRACTED: Revolutionizing medical imaging: A cutting-edge AI framework with vision transformers and perceiver IO for multi-disease diagnosis. *Computational Biology and Chemistry*, *119*, 108586. https://doi.org/10.1016/j.compbiolchem.2025.108586

Lipinski, S., Sadkowski, S., & Chwietczuk, P. (2025). Application of AI in Date Fruit Detection—Performance Analysis of YOLO and Faster R-CNN Models. *Computation*, *13*(6), 149. https://doi.org/10.3390/computation13060149

Lopez-Ruiz, N., Prieto-Sánchez, S., Ramos-Lorente, C. E., Fernández-Moreno, A., Pérez-Ávila, A. J., Álvarez-León, D., Pérez de Vargas-Sansalvador, I. M., García-Cardenas, A., Molina-García, B., Palma, A. J., Capitán-Vallvey, L. F., Martínez-Olmos, A., & Erenas, M. M. (2025). Multimodal sensing approach using remote and onsite data for estimating the pre-harvest ripening stage of Hass Avocado with machine learning algorithms. *Smart Agricultural Technology*, *12*, 101520. https://doi.org/10.1016/j.atech.2025.101520

Ngo, Q. H., Luu, T. H., Nguyen, P. V., El Makrini, I., Vanderborght, B., & Cao, H.-L. (2025). InterDuPa-UAV: A UAV-based dataset for the classification of intercropped durian and papaya trees. *Data in Brief*, *61*, 111843. https://doi.org/10.1016/j.dib.2025.111843

Nguyen, D. T., Bui, T. D., Ngo, T. M., & Ngo, U. Q. (2025). Improving YOLO-Based Plant Disease Detection Using αSILU: A Novel Activation Function for Smart Agriculture. *AgriEngineering*, *7*(9), 271. https://doi.org/10.3390/agriengineering7090271

Nikam, S. B., Lande, S. B., Nagalkar, V. J., Wakchaure, G. C., & Kumar, P. S. (2025). Predictive Classification Model for Quality Grading and Maturity Detection of Dragon Fruit Using Fused Deep CNN Feature and Ensemble Learning. *Journal of Food Processing and Preservation*, *2025*(1), 6938071. https://doi.org/10.1155/jfpp/6938071

Perach, O., Sadeh, R., Avneri, A., Solomon, N., Bonfil, D. J., Ram, O., Greenblatt, H., Lati, R. N., & Herrmann, I. (2025). Harnessing Sentinel-2 imagery and AgERA5 data using Google Earth Engine for developing chickpea mechanistic growth modeling and pre-harvest empirical yield forecast. *Precision Agriculture*, *26*(6), 102. https://doi.org/10.1007/s11119-025-10291-9

Rafi, M. A. S., Senyurek, V., Adeli, A., Yanbo, H., Ball, J. E., & Gurbuz, A. C. (2026). Investigation of machine learning based techniques for crop yield estimation of corn and cotton using multi-sensor data fusion. *Journal of Agriculture and Food Research*, *26*, 102676. https://doi.org/10.1016/j.jafr.2026.102676

Rahman, F. I., Islam, N., & Islam, M. K. (2026). FocalNet: An Attention-Free Novel Architecture for Brain Stroke Classification Using CT Imaging. *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*, 1–6. https://doi.org/10.1109/qpain69676.2026.11545628

Rani, N. S., Bhavya, K. R., Vadivel, A., Vasudev, T., Devadas, R. M., & Hiremani, V. (2025). A Novel Curriculum Learning Training Strategy for Pomegranate Growth Stage Classification Using YOLO Models on Multi-Source Datasets for Precision Agriculture. *IEEE Access*, *13*, 112594–112622. https://doi.org/10.1109/access.2025.3581794

Rayed, M. E., Jim, J. R., Islam, M. J., Mridha, M. F., Kabir, M. M., & Hossen, M. J. (2025). MangoLeafXNet: An Explainable Deep Learning Model for Accurate Mango Leaf Disease Classification. *IEEE Access*, *13*, 93977–94008. https://doi.org/10.1109/access.2025.3571450

Revathi, A. R., & Agash, A. A. (2025). Pomegranate disease diagnosis with severity estimation and treatment remedies using deep learning and RAG-based LLM. *Scientific Reports*, *15*(1), 41278. https://doi.org/10.1038/s41598-025-25160-8

Sar, A., Choudhury, T., Sarkar, T., & Kotecha, K. (2025). Papayafreshnet: a hybrid deep learning framework for non-destructive freshness classification of papayas using convolutional and transformer networks. *Discover Food*, *5*(1), 97. https://doi.org/10.1007/s44187-025-00368-9

Singh, K., Yadav, M., Barak, D., Bansal, S., & Moreira, F. (2025). Machine-Learning-Based Frameworks for Reliable and Sustainable Crop Forecasting. *Sustainability*, *17*(10), 4711. https://doi.org/10.3390/su17104711

Srinivasan, S., Somasundharam, L., Rajendran, S., Singh, V. P., Mathivanan, S. K., & Moorthy, U. (2025). DBAViNet: an effective deep learning framework for fruit disease detection and classification using explainable AI. *BMC Plant Biology*, *25*(1), 965. https://doi.org/10.1186/s12870-025-07015-6

Tapia-Mendez, E., Hernandez-Sandoval, M., Salazar-Colores, S., Cruz-Albarran, I. A., Tovar-Arriaga, S., & Morales-Hernandez, L. A. (2025). A Novel Deep Learning Approach for Precision Agriculture: Quality Detection in Fruits and Vegetables Using Object Detection Models. *Agronomy*, *15*(6), 1307. https://doi.org/10.3390/agronomy15061307

Tipauksorn, P., Luekhong, P., Okada, M., Thongpron, J., Nadee, C., & Yingkayun, K. (2025). Optimized ensemble learning for non-destructive avocado ripeness classification. *Smart Agricultural Technology*, *12*, 101114. https://doi.org/10.1016/j.atech.2025.101114

Wang, Q., Shao, K., Cai, Z., Che, Y., Chen, H., Xiao, S., Wang, R., Liu, Y., Li, B., & Ma, Y. (2025). Prediction of sugar beet yield and quality parameters using Stacked-LSTM model with pre-harvest UAV time series data and meteorological factors. *Artificial Intelligence in Agriculture*, *15*(2), 252–265. https://doi.org/10.1016/j.aiia.2025.02.004

Xie, S., Zhang, N., Liu, L., & Sun, L. (2025). Investigation of the spatio-temporal transferability of large-scale random forest crop classification models and potential influencing factors. *International Journal of Digital Earth*, *18*(2), 2563211. https://doi.org/10.1080/17538947.2025.2563211

Zhu, H., Wang, D., Wei, Y., Wang, P., & Su, M. (2025). YOLOV8-CMS: a high-accuracy deep learning model for automated citrus leaf disease classification and grading. *Plant Methods*, *21*(1), 88. https://doi.org/10.1186/s13007-025-01396-3

Zhuang, H., Zhou, Z., Chen, Y., Yan, Z., Wang, C., Sun, X., Wu, S., Xu, L., Gao, S., Zhang, X., & Du, Z. (2026). Diag-STFN: A diagnostic spatiotemporal multimodal fusion network for global pre-harvest crop yield forecasting. *Ecological Informatics*, *96*, 103860. https://doi.org/10.1016/j.ecoinf.2026.103860